УДК 621.391.037.372

М. Б. Столбов, М. Ю. Татарникова

РАЗДЕЛЕНИЕ РЕЧИ ЦЕЛЕВОГО И СТОРОННИХ ДИКТОРОВ С ИСПОЛЬЗОВАНИЕМ ДВУХМИКРОФОННОЙ СИСТЕМЫ

Рассмотрен метод разделения речи целевого и сторонних дикторов с помощью обработки сигналов двух симметричных микрофонов, ориентированных в противоположных направлениях. Алгоритм обработки обеспечивает пространственное разделение дикторов.

Ключевые слова: *детектор речи, двухканальная фильтрация аудиосигналов.*

Введение. Одной из актуальных практических задач в области обработки речевых сигналов является контроль соблюдения речевого регламента сотрудниками call-центров, диспетчерами (операторами). Данная задача может решаться методами автоматического распознавания речи. На достоверность результатов существенное влияние оказывает окружающая акустическая обстановка, в частности, речь сторонних дикторов, которая может интерпретироваться как речь целевого диктора (РЦД). Для улучшения результатов распознавания необходимо применять специальные методы выделения РЦД.

Для выделения РЦД в сложной акустической обстановке применяются различные системы, включающие в себя как аппаратную, так и алгоритмическую составляющие.

Традиционно для выделения РЦД используются индивидуальные гарнитуры и микрофоны. При этом уровень речи целевого диктора настолько превосходит уровень окружающих помех, что дополнительной обработки сигналов не требуется. Однако в ряде случаев из-за условий работы операторов такой подход не может быть использован.

Также используются микрофонные решетки и алгоритмы пространственной фильтрации [1]. Однако эти методы непригодны для больших помещений и требуют чрезмерных аппаратных затрат.

В последнее время для мониторинга совещаний все большее применение находят распределенные системы микрофонов (claster of microphones), когда каждый диктор снабжается индивидуальным (close-talk) микрофоном [2—5]. Поскольку в таких системах в каждый из индивидуальных микрофонов попадает речь сторонних дикторов, выделение РЦД осуществляется на основе совместной обработки сигналов всех микрофонов. Такие системы требуют оборудования микрофонами всех рабочих мест операторов, при этом предполагается, что сторонние дикторы (операторы) всегда находятся в непосредственной близости к «индивидуальным» микрофонам. Эти ограничения соответствуют сценарию совещания и задаче распознавания речи дикторов (meeting speech recognition) [6], однако не подходят для ситуаций, когда сторонние дикторы меняют местоположение.

Перед авторами стояла задача создания гибкой системы, обеспечивающей выделение РЦД на оборудованных рабочих местах без использования совместной обработки сигналов всех микрофонов. Основная идея предложенного подхода заключается в использовании в рабочей зоне оператора микрофонного блока, состоящего из двух симметричных микрофонов, один из которых направлен в сторону оператора и предназначен для записи его речи, а другой направлен в противоположную сторону и предназначен для приема посторонних звуков.

В работе описан алгоритм совместной обработки сигналов микрофонов, позволяющий выделить речь целевого диктора на фоне акустических шумов окружения и речи сторонних дикторов.

Микрофонный блок. Для выделения речи оператора была предложена схема с двумя противоположно направленными микрофонами. Микрофон основного канала *(m)* предназначался для получения аудиосигнала от оператора, микрофон опорного канала (*r*) — для получения сигнала от акустического окружения. Для избежания проблем калибровки (выравнивания АЧХ микрофонов) и упрощения алгоритмов обработки были использованы симметричные (одинаковые) микрофоны (рис. 1). Для увеличения эффективности пространственного выделения речи оператора и снижения влияния шумов окружения были выбраны суперкардиоидные микрофоны.

а) *б*)

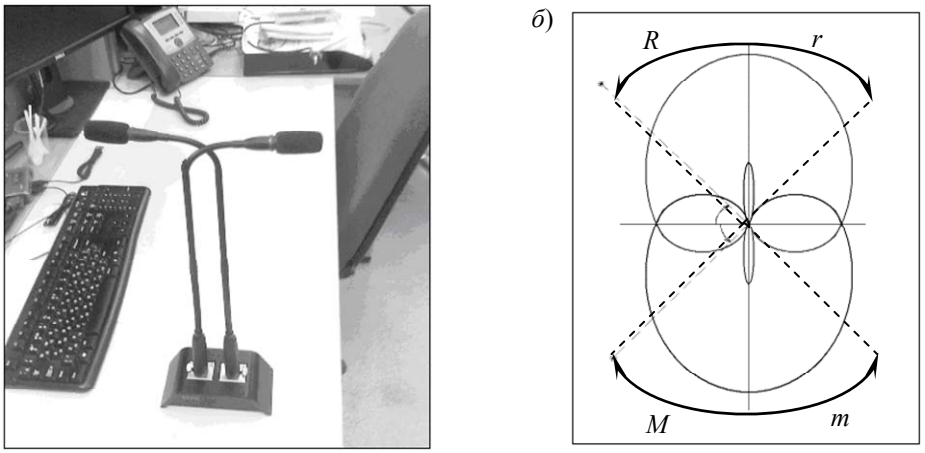

Рис. 1

Выделение речи целевого диктора. Предполагалось, что основным источником помехи является речь стороннего диктора.

Рассматривались два варианта обработки аудиофайла: фильтрация сигналов с целью компенсации речи сторонних дикторов [5, 7] и детектирование участков сторонних дикторов (overlapped speech detection) с целью их исключения из процесса распознавания речи [4—6, 8—11].

Поскольку фильтрация обычно приводит к искажениям РЦД, нами был выбран метод детектирования. Пригодными для дальнейшей обработки и распознавания речи считались сегменты сигнала, содержащие только речь целевого диктора.

Алгоритм выделения РЦД работает следующим образом. На участках речи оператора, не содержащих помех, исходный сигнал основного канала сохраняется, на остальных участках сигнал подавляется на заданную величину (обычно 20 дБ).

Для каждого кадра сигнала микрофона основного канала принималась одна из следующих гипотез:

— H0: паузы речи,

— H1: РЦД,

— H2: речь стороннего диктора,

— H12: одновременная речь целевого и стороннего дикторов.

В работе [11] описан алгоритм сегментации сигнала согласно этим гипотезам на основе оценок критерия отношения мощностей сигналов микрофонов. Предложенный алгоритм обеспечивает пространственную фильтрацию сигналов РЦД и стороннего диктора.

Пример распределения мощностей сигналов основного и опорного каналов (P_r и P_m) для разных участков сигнала приведен на рис. 2 (квадраты — РЦД, кружки и крестики — речь сторонних дикторов).

Эксперименты на записях, выполненных в лабораторных условиях, подтвердили эффективность метода детектирования и алгоритма обработки. Однако в реальных условиях

опытная эксплуатация разработанной системы обнаружила слабые места предложенного подхода.

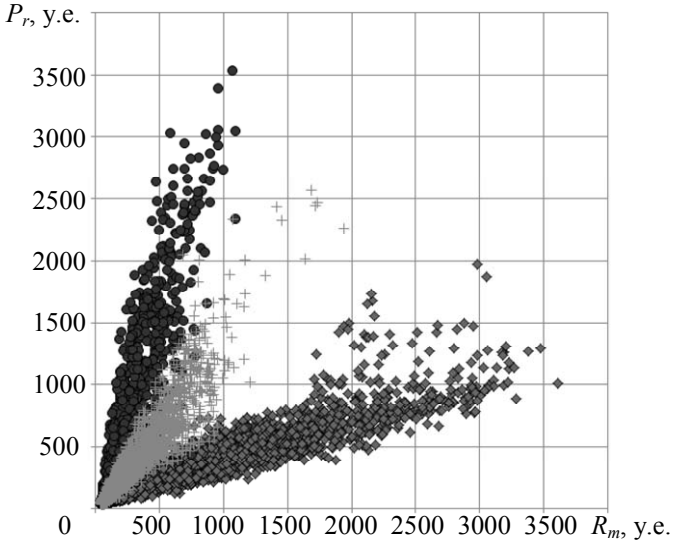

Рис. 2

Выделение РЦД на фоне шумов окружения. Одной из основных проблем выделения РЦД в реальных условиях оказалось большое число источников акустических сигналов вокруг оператора, таких как шум компьютеров, шум принтера, речь сторонних операторов, звук клавиатуры операторов, шелест бумаг, работающие аудиоустройства (динамики, телевизор), наводки и звонки мобильных телефонов и др.

Предложенный ранее критерий, основанный на оценке мощностей основного и опорного каналов, идентифицировал такие звуки как РЦД. Кроме того, речевые сигналы удаленных дикторов поступают в основной и опорный микрофоны приблизительно в равных пропорциях, поэтому критерий отношения мощностей, предложенный в работе [11], оказался неэффективным.

Задача заключалась в выборе новых критериев, позволяющих идентифицировать РЦД в сложной акустической обстановке. Для детектирования участков РЦД на фоне неречевых помех могут быть использованы одноканальные критерии (тональность сигналов, частота пересечения нуля и др.). Однако эти критерии не позволяют выделить РЦД на фоне речевых сигналов (речь удаленных дикторов, речевые сигналы аудиоустройств).

Для выявления акустических событий, порождаемых удаленными источниками, целесообразно использовать так называемые кроссканальные критерии (cross-channel features).

В работах, посвященных детектированию РЦД на фоне речи сторонних дикторов, предложены несколько групп критериев отбора:

— на основе функции когерентности [9],

— на основе функций кросскорреляции [4, 6, 10],

— на основе функций кросс-спектра [8],

— на основе мер подобия амплитудных спектров и спектров мощности [3, 5].

Сегментация сигнала на отдельных кадрах осуществляется по превышению порога одним или несколькими критериями.

Критерии на основе кросс-спектров. По результатам предварительных исследований групп критериев для системы двух микрофонов нами были предложены модифицированные критерии на основе кросс-спектров.

Введем обозначения оценок следующих функций: $\Phi_{mr}(k, t)$ — оценка мгновенного кросс-спектра сигналов основного и опорного каналов, $\langle\Phi_{mr}(k, t)\rangle$ — усредненная по времени оценка кросс-спектра, $\langle\Phi_{mm}(k, t)\rangle$, $\langle\Phi_{rr}(k, t)\rangle$ — усредненная по времени оценка спектров мощности сигналов основного и опорного каналов, k — индекс частоты, t — индекс кадра данных.

На основе этих функций определим интегральные критерии:

$$Q_1 r(t) = \frac{1}{|F|} \sum_k \frac{|< \Phi_{mr}(k,t) >|}{< \Phi_{mm}(k,t) >},$$

$$Q_1 m(t) = \frac{1}{|F|} \sum_k \frac{|< \Phi_{mr}(k,t) >|}{< \Phi_{rr}(k,t) >}.$$

Проверка кросс-спектральных критериев на ряде фонограмм, сделанных в реальных условиях, показала возможность их использования для разделения близких и удаленных источников акустических сигналов. Пример для критериев $Q_1m(t)$, $Q_1r(t)$ приведен на рис. 3.

Видно, что критерий $Q_1m(t)$ позволяет детектировать РЦД (участки 1—3,5; 12—14,5; 20—22 с), $Q_1r(t)$ — речь находящегося вблизи стороннего диктора (6—11,5; 15—15,5 с) в присутствии фонового звука телевизора. На участке звука телевизора в отсутствие речи (участок 3,5—5,5 с) оба критерия оказались нечувствительными к звуку телевизора.

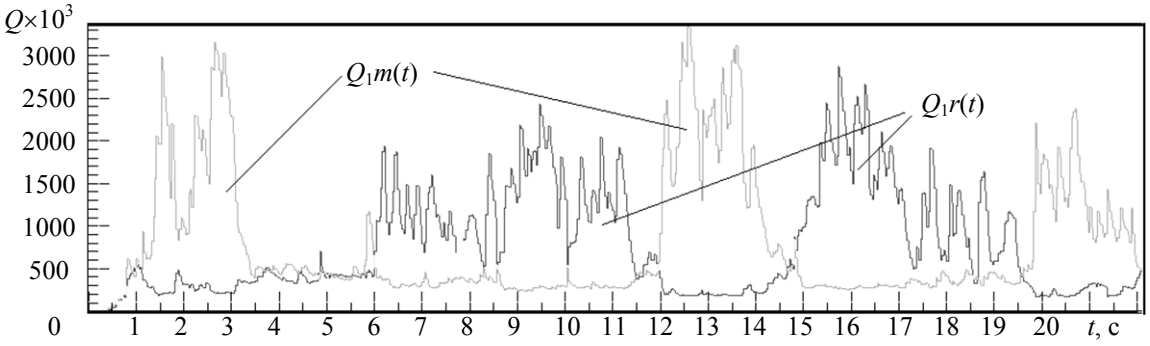

Рис. 3

Исследования на реальных записях показали робастность кросс-спектральных критериев по отношению к акустическим сигналам удаленных источников окружения. Исследование области применения этой группы критериев является предметом дальнейшей работы.

Заключение. Предложен метод выделения участков речи целевого диктора на фоне речи сторонних дикторов и акустических помех. Основная идея предложенного метода заключается в детектировании участков сигнала с речью целевого диктора и исключении участков помех и речи сторонних дикторов.

Предложена система из двух симметричных противоположно направленных суперкардиоидных микрофонов. Алгоритм детектирования речи целевого диктора основан на оценке кроссканальных критериев.

Предложен кросс-спектральный критерий, позволяющий разделить близко расположенные и удаленные источники речевых сигналов. Основным достоинством предложенной системы является простота и применимость в широком диапазоне практических ситуаций, ограничением (по сравнению с использованием гарнитуры) — потеря фрагментов речи целевого диктора на участках присутствия акустических помех и речи близко расположенного стороннего диктора. Более полное исследование предложенного критерия, в частности его применимость для выделения РЦД в сложной акустической обстановке при совместном использовании нескольких критериев, является предметом дальнейшей работы.

Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01).

СПИСОК ЛИТЕРАТУРЫ

1. *Morgan P., George E., Lee T., Kay M.* Co-Channel Speaker Separation // Proc. of the ICASSP. 1995. Vol. 1. P. 828—831.

2. *Nasu Y., Shinoda K., Furui S.* Cross-Channel Spectral Subtraction for meeting speech recognition // Proc. of the IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). 2011. P. 4812—4815.

3. *Xiao B.* et al. Overlapped speech detection using long-term spectro-temporal similarity in stereo-recording // Proc. of the IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). 2011. P. 5216—5219.

4. *Kumatani K.* et al. Channel selection based on multichannel cross-correlation coefficients for distant speech recognition // Proc. of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 2011). UK, 2011. P. 1—6.

5. *Yakoyama R.* et al. Overlapped Speech Detection in Meeting Using Cross-Channel Spectral Subtraction and Spectrum Similarity // Proc. Interspeech. 2012. P. 1—4.

6. *Boakye K., Stolcke A.* Improved Speech Activity Detection Using Cross-Channel Features for Recognition of Multiparty Meetings // Proc. Interspeech. USA, 2006. P. 1962—1965.

7. *Cao Y., Sridman S., Moody M.* Multichannel Speech Separation by Eigendecomposition and Its Application to Co-Talker Interference Removal // IEEE Trans. on SAP. 1997. Vol. 5, N 3. P. 209—219.

8. *Wrigley S. N., Brown J., Wan V., Renals S.* Speech and Crosstalk Detection in Multichannel Audio // IEEE Trans. on SAP. 2005. Vol. 13, N 1. P. 84—91.

9. *Yen K.-C., Zhao Y.* Robust Automatic Speech Recognition using a multi-channel signal separation front-end // Proc. Fourth Intern. Conf. on Spoken Language, ICSLP. 1996. Vol. 3. P. 1337—1340.

10. *Laskowski K., Schulttz T.* A geometric interpretation of non-target-normalized maximum cross-channel correlation for vocal activity detection in meetings // Proc. of NAACL HLT. 2007. P. 89—92.

11. *Stolbov M., Tatarnikova M.* Speech and Crosstalk Detection for Robust Speech Recognition Using a Dual Microphone System // Proc. of 15th Intern. Conf. on Speech and Computer, SPECOM. 2013. P. 310—318.

Сведения об авторах

Михаил Борисович Столбов — канд. техн. наук, доцент; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; ООО „ЦРТ-инновации“, Санкт-Петербург; старший научный сотрудник; E-mail: stolbov@speechpro.com

Марина Юрьевна Татарникова — ООО „ЦРТ-инновации“, Санкт-Петербург; старший научный сотрудник; E-mail: tatmar@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.